DOCUMENT RESUME

ED 238 263 FL 014 124

AUTHOR Carr, Marion
TITLE A Five-Step Evaluation of a Holistic Essay-Evaluation Process.
PUB DATE 22 Jul 83
NOTE 19p.; Paper presented at the summer TESOL meeting (5th, Toronto, Ontario, July 21-23, 1983).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Admission Criteria; *College Applicants; *English (Second Language); *Essays; Evaluation Methods; Higher Education; *Holistic Approach; Second Language Instruction; *Student Evaluation; Test Reliability

ABSTRACT

The faculty of an intensive program of English as a second language for college-bound students, questioning the objectivity of faculty evaluations of non-native college applicants' written essays, assessed the existing evaluation process, reformed it, tested it, and planned for ongoing development. In the first stage, readers read and graded essays according to practices currently in use, approaching the essays holistically, without breakdown of elements, and students wrote on a choice of topics. The lack of precision in grading was examined. The second stage involved devising new evaluation criteria. Faculty at five levels of English instruction analyzed instructional program goals and objectives, developed criteria from them, and applied them to the previously-graded essays. In the third stage, the revised criteria were tested in a midterm examination and the essays assigned randomly to the faculty for grading. The new evaluation process resulted in individual grids of writing strengths and weaknesses to be presented to the students. The fourth stage, assessment of the new criteria's usefulness, showed that the faculty valued these elements of the process; reduced evaluation time, a more humanistic testing procedure, helpful faculty discussion, more professional logistics, and a spirit of cooperation. Negative comments focused on the format of the grid of student strengths and weaknesses, the time involved in recording grades, and the use made of the results. The fifth stage involves an ongoing examination of the criteria's usefulness and reliability. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED238263

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARION CARR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Five-Step Evaluation of a Holistic Essay-Evaluation Process

by

Marion Carr PBVM, Ed.D.

presented at

1983 Summer TESOL Institute

July 22,1983

FL 014 124

A Five-Step Evaluation of a Holistic Essay-Evaluation Process*

Marion Carr

Evaluating essays and compositions has always been challenging to English instructors on any level. The evaluation process takes on added significance when applied to the essays and compositions of non-native English speakers seeking admission into American colleges and universities. "Project: Essay Evaluation," which is described in the following article proceeded through five separate but sequential steps from questioning the existing essay evaluation process to rewriting it, piloting the new version, reviewing the results, and planning for on-going development of the process. The article includes statistical reports on the convergence of reader assessments and on intra-reader reliability.

How objective can essay evaluation be? The faculty of an intensive program of English as a second language for college-bound students

---

*This paper was presented at the 1983 Summer TESOL Institute, Toronto, Canada.

Dr. M. Carr, assistant professor at the World English Center of the University of San Francisco (California) was coordinator of "Project: Essay Evaluation."

questioned this. Praxis revealed that a single essay read by five different evaluators could, on a scale of 100, have a variation of as many as 15 points. In reality, were this evaluation the sole criterion for a student's grade placement in an ESL program or college admission's "writing" clearance, there could be a difference of three ESL grade levels or admissions/non-admissions, depending upon the intra-reader reliability.

Eighteen months ago this faculty under the leadership of the curriculum coordinator undertook a project called "Project: Essay Evaluation." The following five steps of the project report the direction, the findings, and the application of the results to the evaluation of student essays.

Step One: Holistic Grading or Not?

The faculty first reviewed the current practice:

1. Essays were read for placement purposes/"writing" clearance for college admissions;
2. Essays were assigned percentage points in order to be averaged with other standardized test scores;
3. Percentage points for the essays were assigned on the basis of descriptors for level placement or college admissions;
4. Each essay was evaluated by a minimum of three readers;
5. Evaluators' scores on a single essay had to fall within a 10-point maximum spread or the essay had to be read by additional readers until this condition was met;

6. The type of point assessment was holistic, that is, there was no break-down of points according to various elements within a composition;
7. Evaluation-orientation consisted of reviewing the descriptors, reading sample essays which had been assigned a common grade by three evaluators, and then practicing grade-assignment with other previously graded compositions and comparing results;
8. Evaluators were selected from the intensive English faculty but not restricted to the writing and grammar instructors;
9. Topics for the essays were assigned, but students could choose one out of two or one out of three;
10. Students were given thirty minutes for the entire writing process to be accomplished without use of either practice paper or dictionaries.

As the discussion of the existing holistic grading system continued, the participants surfaced distinct advantages, in line with other proponents of holistic grading (Cooper, 1977; Oller, 1979). The faculty saw the composing process and essay writing as non-formulaic. Essay writing, by its very essence, is the communication of a central idea or ideas to a specific "audience" by an individual through clarification, restatement, illustration, expansion, and a variety of other techniques and literary devices. Can an evaluator mathematically add up these varying elements--so diversified from one essay to another? Holistic evaluation seemed to safeguard the essence of the art form, allowing for individual strengths and weaknesses, presence or absence of character-

istics of the art in the essays of ESL students.

Since the decision to retain holistic grading was consensual, what other aspects of the evaluation process were to be considered? The criteria--as found in the existing descriptors--were then scrutinized. These lacked preciseness; two different grades could be assigned to the same essay on the basis of the descriptors as the following illustrate:

(A) Poor everything but sense can be made of it. A lot of evidence of the writer's native language. Mostly elementary structures with very few complex sentences. Several incomplete sentences. Most sentences have at least one error. In some cases, rather accurate but extremely short with only simple structures. About 1 page or less.

(B) Very simple. No sentence complexity or only one or two botched attempts at complex sentences. Spelling and mechanics, very elementary. Some apparent translation. Poor vocabulary. 1 page or less.

When the elements in the descriptors are analyzed the lack of precision becomes apparent:

Sentence structures	(A)	"mostly elementary structures with very few complex sentences"
	(B)	"no sentence complexity or only one or two botched attempts at complex sentences"
Translation	(A)	"evidence of the writer's native language"
	(B)	"some apparent translation"

Length (A) "about 1 page or less"

(B) "1 page or less"

With a ten-point spread between these two descriptors, it was easy to place the lack of precision in the evaluators' assessments. Here was the area to start to tighten the evaluation process.

Step Two: What Are the Criteria?

The process for establishing new criteria--or better restating of the underlying criteria in the existing descriptors--developed into a tripartite discussion with each part requiring a minimum of two hours, exclusive of any preliminary study, reading, reviewing of current criteria, or re-thinking the system.

The first meeting engaged the writing and grammar instructors on each of the five intensive English levels. These instrucors studied the goals and objectives for their respective levels in the areas of grammar and writing. In the preceding year, a five-level, five-skill syllabus had been published in-house; therefore, the goals and objectives in it were readily available to the instructors. Although some of the current level coordinators and some of the instructors had participated in the syllabus development, there were sufficient newcomers to make this phase of the study necessary. To prepare for the second meeting for establishing criteria, the grammar and writing instructors analyzed the goals and objectives under these headings: grammar/punctuation, diction/spelling; content/organization; and structural complexity/diversity.

In the second meeting, level coordinators and instructors wrote criteria derived from the goals and objectives under the major headings. Following the level meetings, a general meeting was held to test the sequence of the criteria which were posted by levels on newsprint side by side. Discussion ensued wherever there was overlap, and the terms "introductory," "adequate," and "mastery" (which appeared in the syllabus) were suggested to distinguish the phases of expectation of similar techniques and skills in the students' writing. In the first level, structural complexity would be introductory, but at the fifth level, it would be mastery.

One of the most enlightening aspects of the second step was the discovery of differences in the priorities of the four major categories. Some wanted to emphasize grammar as key to assessing an ESL student's writing skill; others stressed the content. Since the primary purpose of the writing sample was the assessment of the ESL student's ability to write a composition or essay in English, the group finally reached the conclusion that structural complexity /diversity was a prime--if not the prime--indicator of written language fluency. Examples were cited from ESL students' compositions in which there were simple sentences in short paragraphs with few errors and a very limited vocabulary (A), in contrast to the writing from an ESL student on the same level in which there was more native "flow" to the writing, despite the errors (B):

(A) My parents visited in United States of America two years ago.

They're astonished and liked America. My father wrote to me. He thought the America was very beautiful, clean, and wide.

(B) Im learning English because I need it for my majo· in Mexico. although I don't want to do my major in U.S.A. the English is so important in Mexico too, because almost all the books in my major are in English.

The third meeting in the criteria phase of the project was to apply the newly-derived criteria to a set of previously graded essays and test the correlation. The criteria seemed to function satisfactorily so the project moved into step three.

Step Three: How Effective Is the Pilot Study?

The revised criteria were tested in a midterm examination framework. The instructors who had worked on the criteria--the entire faculty of the intensive program--became the evaluators of the midterm essays. Every effort was made to involve all instructors in the evaluation process, whether the individual member was a writing instructor or not. The conditions of the midterm were identical to the previous placement testing in all aspects but two. A thirty-minute writing sample on an assigned topic was evaluated by three readers holistically. The resulting grade to be valid had to fall within the ten-point spread. The two conditions which varied from the previous testing of writing ability were 1. restricting the sample to a single topic in order that con-

centration was on the students' skills as evaluated according to the new criteria, rather than on the students' knowledge of one or other of the topics; and 2. the students were given a copy and an explanation of the criteria in their respective writing classes one week before the essay examination. During this week students did sample writings and the writing instructors demonstrated with these samples how the criteria applied.

On the day of the essay examination, all faculty members as evaluators were assigned to small groups under the five level coordinators. Each group was given a random batch of the midterm essays to evaluate. Each reader graded the essay holistically and then designated on a grid specific strengths and weaknesses under the four major categories (listed in Step Two). One student's writing, for example, might receive a score of 60 with strengths in vocabulary and weaknesses in spelling noted on the grid; another might receive the same holistic score of 60 but with strengths in grammatical accuracy. If the three evaluators of an essay did not submit scores within the 10-point spread, they discussed among themselves the basis for their evaluations, and a fourth reader was called in. Through the discussions, the overall perspective of the criteria and the priorities set by the group were in focus for the evaluators. After each 50-minute evaluation period (with a 10-minute break) the evaluators, with the exception of the level coordinators, rotated to a new group to repeat the process with another random batch of essays. The entire evaluation process, including registering the scores of the 675 readings, was accomplished in three hours and forty minutes. Of the total readings, there was a convergence (the exact same score)given

in 58.22 percent of the readings and a divergence in 41.78 percent. Among the divergent scores, 27.11 percent had a five-point spread; 10.22 percent, a 10-point spread; and 4.44 percent, a 15-point spread.

Step Four: After the Pilot, What Changes?

The week following the midterm, the students were given their midterm essays, the individual averaged holistic scores, and their respective grids of strengths and weaknesses. The writing instructors used different techniques in reviewing the above data with their students. One advanced level writing teacher reassigned the topic, allowing the students as much time in the fifty-minute class hour as they wanted to write, copy read, and revise. Students in this class then compared their midterm version with the second in-class version and noted their own lack of copy reading in the midterm which resulted in avoidable errors. All writing instructors, no matter what approach they took reported positive receptivity of the evaluations by the students and a renewed interest in the writing process.

Positive responses from the evaluators included the following: there was 1. a great time reduction in the overall evaluation process; 2. a more humanistic approach to the evaluation process that designated a number of essay readings (12-15 for each reader) followed by a break; 3. helpful discussions among the three evaluators that kept the holistic perspective when there was extreme divergence in scores; 4. the logistics--separate rooms for each team of evaluators--which added a more professional atmosphere; and 5. the spirit of cooperation because of the total faculty involvement in the evaluative process.

On the other hand, the negative elements that were expressed centered on the grid for strengths and weaknesses. Those that responded to it negatively saw positive feedback to the students if there had been a composite grid made out by the three evaluators and not three separate grids; however, they felt the time required to do this was excessive in the evaluative process for placement and admissions. Another negative element--one that had not entered into the original planning--was the time required to record the scores for the 675 readings. And finally, there were some instructors who wanted a more formalized plan for handling the post-midterm data in the writing classes.

Besides the positive and negative comments on the pilot study, individual suggestions were expressed. A pre-writing stage in the test, in accord with the instructional emphasis on pre-writing, was suggested. This resulted in the inclusion of a five-minute period for thinking, outlining, and/or note-taking on subsequent essay examinations. Now, on every written examination the following is printed:

> DIRECTIONS: In this test you will have five (5) minutes to plan and thirty (30) minutes to write....
>
> REMEMBER: First plan, then write, then correct.
>
> This space is for your notes. Your notes will NOT be graded.
>
> ---
>
> (Space for Notes)
>
> ---

Students were informed that they could plan or make notes in either English or their native language. A sample of an introductory level student from Mexico (whose paper was previously quoted) follows:

---

1. Yo estudio ingles porque lo necesito
   - para la escuela
   - los libros penso hacer
   - mi mayor aqui
2. Ingles es la lengua universal
   - a todo lado en que vaya lo usaré
   - la estudio aqui porque es el ingles mas usado
3. Me gusta hablar ingles es un idioma bonito

Another positive outcome of the pilot study was the distribution of copies of the criteria to all students before the midterm. And a third outcome was the revival in the intensive program of the "Topic Bank." For this each instructor was asked to submit between five to ten essay topic suggestions so that there were over 200 topics generated for future in-class writing assignments. Although these were not the topics for final or midterm examinations, they gave students wide exposure to the variety of subjects for composition and essay writing.

With regard to the negative reaction to the grid, mostly on the basis of its time consumption, a decision was made to drop this element in the evaluative process.

Step Five: Repeat Measure, What Did It Produce?

For the past year, the criteria and evaluation process have been applied to the essay evaluation during midterm and final examinations in the fall, intersession, and spring semesters. A comparison of the

results of the pilot study with the midterm and final examination statistics of Spring '83 reveals great consistency and a forward movement toward better intra-reader reliability for these more than 600 essays or more than 1800 evaluations. (See Table 1.)

Also, in Spring '83, it has been possible to compare the separate TOEFL scores for grammar and structure, reading and vocabulary, and listening comprehension with the writing sample assessments made by the intensive program evaluators. The comparison shows high correlation and high reliability coefficients for the 14 classes. (See Table 2.)

A final comparison in this evaluation of the holistic essay-evaluation process compared the results of the "teams" of evaluators, made up of the level coordinators and the rotating faculty/evaluators. A sample of this comparison shows that Team A evaluated 86.24 percent with a 5-point differential; Team B, 84.25 percent; Team C, 76.0 percent; and Team D, 85.15 percent. It must be remembered that with a five - point difference, the student's average score varies only three points; and with a ten - point difference, the student's average varies from 3 to 5 points. (See Table 3.)

Conclusion: Is This the End?

In no way is the above process the answer to making the evaluation of essays fool-proof. This evaluation process is but one model that has worked successfully in tightening the criteria for a group of evaluators and in informing the students of the basis for their essay evaluations. Its continued success will depend upon how thoroughly newcomers to the

current group of evaluators "ingest" the criteria and participate in the process of developing greater intra-reader reliability. It is anticipated that continued study of this process will also produce new and improved ways of looking at essay evaluation.

COMPARISON OF SCORING

PILOT '82, Mid-Term '83, Final '83

		Convergence 5 pt.	10 pt.	15 pt.
PILOT '82	58.22	27.11	10.22	4.44
	85.33			
MID-TERM '83	57.01	27.71	13.07	1.73
	84.82			
FINAL '83	56.90	31.45	9.98	1.66
	88.35			

TABLE 1.0

COMPARISON OF TOEFL SCORES WITH ESSAY SCORES

Essay score with	TOEFL STRUCTURE	LISTENING	READING
r	.96	.95	.92
R^2	92.16	90.25	84.0
validity	99.9	99.9	99.9

TABLE 2.0

INTRA-READER RELIABILITY BY TEAMS

Convergence	
Team A	60.55
B	43.25
C	40.0
D	38.09
Within 5 points	
A	25.69
B	41.02
C	36.0
D	47.6
Within 10 points	
A	11.93
B	15.38
C	20.0
D	14.28
Within 15 points	
A	0.009
B	0.0
C	4.0
D	0.0

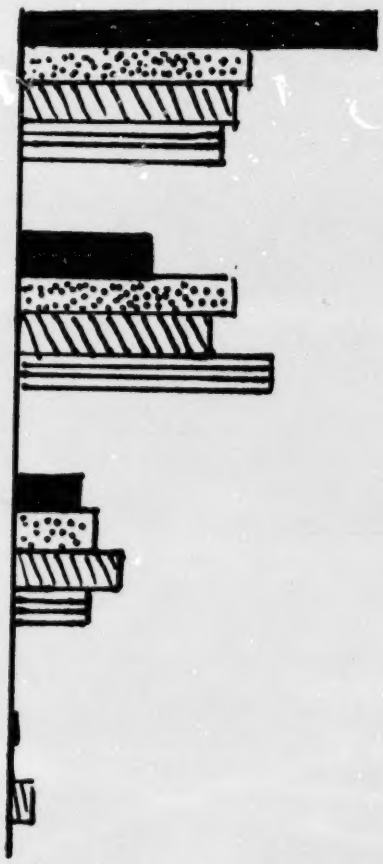

TABLE 3.0

REFERENCES

Cooper, C.R. and L. Odell (eds.). 1977. Evaluating writing: describing, judging, measuring. Urbana, Ill. : National Council of Teachers of English.

Jacobs, Holly L. and others. 1981. Testing ESL Composition. Rowley, Mass.: Newbury House.

Oller, J.W. Jr. 1979b. Language Tests at School. London: Longman Group Limited.